

Estimated P-values in Discrete Models: Asymptotic and non-asymptotic effects

BY CHRIS J. LLOYD

Melbourne Business School, Carlton, 3053, AUSTRALIA
c.lloyd@mbs.edu

SUMMARY

The exact null distribution of a P-value typically depends on nuisance parameters unspecified under the null. For discrete models and standard approximate P-values, this dependence can be quite strong. The estimated (or bootstrap) P-value is the exact probability of the P-value being no larger than its observed value, with the null estimate of the nuisance parameter substituted. For continuous models, it is known that such ‘bootstrap’ P-values deviate from uniformity by terms of $O(m^{-3/2})$, where m is a measure of sample size. The main difficulty with discrete models is the breakdown of asymptotics near the boundary. The aim of this paper is to numerically examine the accuracy of standard and bootstrap P-values for discrete models. We examine a range of binomial models, test statistics and look at testing both canonical and non-canonical parameters. When departures from uniformity are averaged across the nuisance parameter, we find that errors of bootstrap P-values appear to improve at rate $O(m^{-1})$. When interest lies in the maximum error, we find that errors hardly seem to decrease with sample size. Nevertheless, bootstrap P-values enjoy accuracy an order of magnitude better than standard P-values, even for small sample sizes. The reasons why bootstrap works so well in this regard has nothing to do with asymptotics and it is explained how bootstrap can automatically correct small departures from the likelihood based ordering of the sample space.

Some Key Words: nuisance parameters; exact test; tests of independence; r-star; bootstrap

1 Introduction

Suppose we have a model $\pi(y; \psi, \lambda)$ for discrete data $Y \in \mathcal{Y}$ that depends on (ψ, λ) . We are specifically interested in testing a null value ψ_0 by computing null tail probabilities of some test statistic T . We denote all probabilities simply by $\pi(\cdot; \lambda)$, suppressing ψ_0

where there is no confusion. Consider the tail probability

$$\Pi(t, \lambda) := \Pr(T(Y) \geq t; \psi_0, \lambda) = \sum_{y: T(y) \geq t} \pi(y; \lambda)$$

where $t = T(y)$. Since this depends on λ which is unspecified under the null, an obvious route is to replace λ by an estimate, giving $\Pi(t, \hat{\lambda}) = \sum_{y: T(y) \geq t} \pi(y; \hat{\lambda})$ which is an estimate of the true P-value. We are interested in the order of error that is incurred when this estimated P-value is quoted as if it were a true P-value.

We will suppose that the estimate $\hat{\lambda}$ differs from λ by $O_p(m^{-1/2})$ where m is some measure of sample size or information. Supposing only that the model $\pi(\cdot; \lambda)$ is smooth in λ , it follows that $\Pi(t, \lambda)$ is a smooth function of λ . Therefore, in the most general case, $\Pi(t, \hat{\lambda})$ differs from $\Pi(t, \lambda)$ by $O(m^{-1/2})$. However, in hypothesis testing context, T is not an arbitrary statistic; it is asymptotically pivotal (usually normal) under the null which means that $\Pi(t, \hat{\lambda})$ is not only smooth in λ but depends on λ less as sample size increases. Suppose there is an expansion of $\Pi(t, \lambda)$ in powers of $m^{-1/2}$ of the form

$$\Pi(t, \lambda) = 1 - \Phi(t) + e_1(t, \lambda)m^{-1/2} + O(m^{-1})$$

where $e_1(t, \lambda)$ is a smooth function of λ and $O(1)$. It follows that

$$\Pi(t, \hat{\lambda}) - \Pi(t, \lambda) = (e_1(t, \hat{\lambda}) - e_1(t, \lambda))m^{-1/2} + O(m^{-1})$$

and since $e_1(t, \hat{\lambda}) - e_1(t, \lambda) = O(m^{-1/2})$ it follows that $\Pi(t, \hat{\lambda})$ differs from $\Pi(t, \lambda)$ by $O(m^{-1})$. This is a form of parametric bootstrap. There are various ways that the bootstrap P-value can be implemented depending on where the estimator is substituted: into the error term $e_1(t, \lambda)$, into an expression for the mean and variance of T which is then used to create a more accurate pivotal or directly into $\Pi(t, \lambda)$ as we are investigating here.

The P-value under study in this paper involves replacing λ not by an arbitrary estimator of λ but by the ML estimator $\hat{\lambda}_0$ under the null. Since this restricted estimator also differs from λ by $O_p(m^{-1/2})$, it would seem that nothing is gained, at least in terms of asymptotic error rates. For continuous models, it has been shown by various authors, most generally by Lee and Young (2005), that inferential errors are reduced to $O(m^{-3/2})$. The prime requirement is that the test statistic T is a standard likelihood based one and that the estimator $\hat{\lambda}_0$ is the standard restricted ML estimator. It can be shown that $\hat{\lambda}_0$ is asymptotically uncorrelated with T , which results in the $O(m^{-1})$ terms vanishing. This is not true of the unrestricted ML estimator. It is not clear what the error rate will be for discrete models.

2 Background on asymptotic theory

The argument above that bootstrap P-values have error $O(m^{-1})$ is due to Beran (1988). He also introduced a useful way of re-expressing the problem in terms of adjustments to P-values that improve their uniform distribution rather, than as approximations to tail probabilities of a test statistic. Rather than focus on the event $\{T(Y) \geq T(y)\}$ we consider approximately uniformly distribution P-values $P(Y)$, often of the form $1 - \Phi(T(Y))$ for some test statistic T . Then

$$\Pi(t; \lambda) = \Pr(T(Y) \geq t; \psi_0, \lambda) = \Pr(P(y) \leq p; \psi_0, \lambda)$$

The benchmark for a P-value is that the right hand side should equal p for all p in the support set $P(\mathcal{Y})$. His bootstrap P-value is $\Pi(t; \hat{\lambda})$. Since this is an estimate of the true significance, we will henceforth denote it $\hat{P}(Y)$. Assuming that the distribution of T is continuous, Beran (1988) showed that the distribution of this approximate pivot differs uniform by terms and of $O(m^{-(d+1)/2})$ where the initial approximate P-value $P(Y)$ differs by terms of order $m^{-d/2}$. For standard one-sided tests based on the normal approximation, $d = 1$. His arguments also showed that further application of this adjustment, which he called pre-pivoting, results in further half power order of error improvements. It is unclear if these results apply in the discrete case (see comments after Beran's equation 3.2) and he did not distinguish between estimators of λ .

The earliest result concerning the accuracy of substituting $\hat{\lambda}_0$ rather than $\hat{\lambda}$ is DiCiccio and Stern (1994) who showed that when $T = \tilde{R}$, a bias adjusted version of the signed root LR statistic R , normalising by null estimates of the mean and standard deviation results in distributional error of $O(m^{-3/2})$ for continuous models. This work seems to be motivated as an extension of well-known Bartlett adjustment to the LR statistic. DiCiccio, Martin and Stern (2001) extended the result to $T = R$, the SRLR without bias adjustment. They also show a second result concerning $\Pi(t, \hat{\lambda}_0)$ and its relation to $\Pi(t, \lambda)$. Specifically, in the middle of p70 ,their main claim is that

$$\Pr(T(Y) \geq t_{\text{obs}}; \psi_0, \hat{\lambda}_0) - \Pr(T(Y) \geq t_{\text{obs}}; \psi_0, \lambda) = O\left(m^{-3/2}\right). \quad (1)$$

There are two problems with this formulation, one conceptual and one technical. First, it is not clear that this is a useful property because the LR statistic T may have erratic properties in discrete models. In this case, approximating the tail probabilities of T may not be a sensible aim. Rather, the more pertinent issue is the statistical properties of the estimated tail probability as a P-value in, and of, itself.

Second, the property as stated is ill-defined. The left hand side is random through $\hat{\lambda}_0$ and so the O term should apparently be O_p . More confusing is the role of t_{obs} in this equation. When we change sample size then the support of T changes and so t_{obs} will no longer be an observable value of T . So it makes more sense to interpret this equation for any fixed value t within the normal deviation region. The claim then is that for the LR statistic $\Pi(t, \hat{\lambda}_0) - \Pi(t, \lambda)$ is $O_p(m^{-3/2})$.

To measure this property, denote the sample space by \mathcal{Y} with cardinality N and the set $\{y : T_m(y) \geq t\} \subset \mathcal{Y}_m$ by $R_m(t)$. The absolute value of the l.h.s of (1) is

$$\begin{aligned} \epsilon(y_j, \lambda, t) &:= |\Pr(T_m \geq t; \psi_0, \hat{\lambda}_0(y_j)) - \Pr(T_m \geq t; \psi_0, \lambda)| \\ &= \left| \sum_{y \in R_m(t)} \pi(y; \hat{\lambda}_0(y_j)) - \sum_{y \in R_m(t)} \pi(y; \lambda) \right| \\ &= \left| \sum_{y \in R_m(t)} (\pi(y; \hat{\lambda}_0(y_j)) - \pi(y; \lambda)) \right| \end{aligned}$$

for each y_j in the sample space. We then calculate some measure of the size of $\epsilon(Y_j, \lambda, t)$ with respect to the probability distribution $\pi(y_j; \lambda)$. For instance the mean is

$$\bar{\epsilon}(\lambda, t) = \sum_{y_j \in \mathcal{Y}_m} \pi(y_j; \lambda) \left| \sum_{y \in R_m(t)} (\pi(y; \hat{\lambda}_0(y_j)) - \pi(y; \lambda)) \right| \quad (2)$$

Computing the bootstrap P-value itself required $O(N)$ computations but computing its accuracy requires $O(N^2)$ computations, which is very difficult to compute for reasonable sample sizes.

Lee and Young (2005) further extended the theory to any test statistic T which has asymptotic form

$$T = \frac{V_{\psi\psi}U_{\psi} + V_{\psi\lambda}U_{\lambda}}{V_{\psi\psi}^{1/2}} + O_p(m^{-1/2})$$

where U is the score function and V is the inverse of the Fisher information matrix. This includes the standard likelihood based test statistics. For instance, substituting $(\psi_0, \hat{\lambda}_0)$ into the first term gives the statistic $\hat{V}_{\psi\psi}^{1/2}U_{\psi}$ evaluated at $(\psi_0, \hat{\lambda}_0)$ which is the score statistic.

Their main claim concerns the bootstrap P-value $\Pi(t(y), \hat{\lambda}_0(y))$ treated as a random variable. Notice that this random variable depends on the data through $T(Y)$ and $\hat{\lambda}_0(Y)$. In particular, it induces a different ordering on the sample space than does the original approximate pivotal $1 - \Phi(T)$. Denoting the bootstrap P-value by $\hat{P}(Y)$, they state that for any $\alpha \in [0, 1]$

$$\Pr(\hat{P}(Y) \leq \alpha; \psi_0, \lambda) = \alpha + O(m^{-3/2}) \quad (3)$$

for any λ for which an Edgeworth expansion is valid. This property is directly relevant to the performance of $\hat{P}(Y)$ as a P-value, but does not say anything directly about how well it might approximate the original tail probability.

Their argument is based on an expansion

$$\Pi(t, \lambda) = 1 - \Phi(t) + \frac{e_1(t, \lambda)}{m^{1/2}}\phi(t) + \frac{e_2(t, \lambda)}{m}\phi(t) + O(m^{-3/2})$$

and also approximating

$$e_j(t, \hat{\lambda}) = e_j(t, \lambda) + (\hat{\lambda} - \lambda)e'_j(t, \lambda) + O(m^{-1}),$$

so that

$$\Pi(t, \hat{\lambda}) = \Pi(t, \lambda) + \frac{m^{1/2}(\hat{\lambda} - \lambda)e'_1(t, \lambda)}{m}\phi(t) + O(m^{-3/2})$$

They show that the covariance of $m^{1/2}(\hat{\lambda} - \lambda)$ and T vanished to first order, so that the $O(m^{-1})$ term in the distribution of $\hat{P}(Y)$ vanishes. The key property of the restricted ML estimator that leads to higher accuracy is that deviations of $(\psi_0, \hat{\lambda}_0)$ from (ψ_0, λ) under the null are, to first order, proportional to $(0, U_\lambda)$ and that U_λ is uncorrelated with T above. It is intuitively obvious then that using the unrestricted estimator $(\hat{\psi}, \hat{\lambda})$ runs the risk of being contaminated by deviations of $\hat{\psi}$ from ψ_0 which are correlated with the test statistic itself.

The numerical illustrations of Lee and Young are all from continuous models and their conclusion about using the null ML estimator $\hat{\lambda}_0$ rather than the non-null estimator is that "in practice the gains are typically rather slight." It is not clear what error rates are being claimed for discrete data. To compute the left hand side of (3) we must determine the set $R_\alpha = \{y : \hat{P}(y) \leq \alpha\}$. This requires computing all possible values of $\hat{P}(y)$ which is typically an $O(N^2)$ computation.

Most recently, DiCiccio and Young (2008) show that, for canonical parameters of exponential families, the estimated P-value is not only third order accurate but also approximates the exact conditional P-value to the same order. This is a remarkable result and means that for full exponential families, substitution of a sufficient estimate of the nuisance parameters is almost equivalent to conditioning on these same sufficient statistics. This theory is claimed to apply also to discrete distributions where the estimated P-value provides an approximation to the mid-P from the relevant conditional distribution. Their numerical illustrations suggest however that the error is $O(m^{-1})$, which is the same order of error as achieved by competing analytic methods such as r^* , see Brazzale, Davison and Reid (2007).

For continuous models, it is claimed in Fraser and Rousseau (2008) that replacing λ by the unconstrained $\hat{\lambda}$ gives a P-value that has error $O(m^{-1})$ and that a second application gives error $O(m^{-3/2})$. The result is based on extremely general theory known as high order asymptotics, and requires that the model be resolved into the estimate and a general but unspecified third order ancillary statistic.

2.1 Discreteness

It is worth thinking about how discreteness *per se* may limit the accuracy attainable. It is easy to check numerically that the normal approximation to the binomial has error $O(m^{-1/2})$ and that the mid-P has the same order of error except for the case $p = 0.5$. In this case the skewness is zero and the error is $O(m^{-1})$ and the bias is $O(m^{-3/2})$. For the generalised hypergeometric distribution, the case of equal sample size and independence also involves a symmetric distribution and may prove to be a special case. But as one goes to higher dimensions the discreteness becomes less important

Denote the cardinality of the sample space by N and consider binomial data (y_1, y_2, \dots, y_d) with denominators $m(n_1, n_2, \dots, n_d)$. Then $N = \prod(mn_i + 1) = O(m^d)$. We expect then that the support of a test statistic $T(Y)$ will increase at the same rate, providing that there are not too many ties. On the other hand, the normal deviation region of T , say $\{y : T(y) \leq 2\}$ will asymptotically contain a proportion $m^{-1/2}$ of the sample space. Therefore we would expect that the average distance between the support points within a normal deviation range should decrease like $1/m^{d-1/2}$. This has been confirmed numerically in the table below. The test statistic T was the SRLR statistic for testing zero slope in a logistic regression but with weighted least squares estimates inserted instead of maximum likelihood. The number of sample points in the set $\{y : T(y) \leq 2\}$ was calculated for sample sizes $m(n_1, \dots, n_d)$ with $m = 1, \dots, 10$, Assuming that the cardinality of this set is asymptotically of the form Am^γ the exponent γ was estimated using standard methods.

Table 1: **Asymptotic cardinality of set $\{y : T(y) \leq 2\}$.**

d	2	2	2	3	3	4
(n_1, \dots, n_d)	(5,5)	(4,6)	(2,8)	(5,6,7)	(2,8,3)	(3,4,5,6)
$\hat{\gamma}$	1.52	1.51	1.50	2.51	2.48	3.47

This suggests that, even for a simple bivariate binomial model such as a 2×2 table, gaps in the support of T will be $O(m^{3/2})$ in the normal deviation range and so there is no reason why discreteness of the test statistic itself will prevent P-values that are accurate to the order $O(m^{-3/2})$. Another quite different problem is the behaviour of the test statistic near the boundary of the parameter space. For standard discrete models there is typically a complete breakdown at the boundaries of the parameter space. So to emulate the continuous results one would need to bound the parameter away from the boundary and to modify the E-step accordingly.

3 Measuring the accuracy of a P-value

Let $P_m(Y)$ denote an arbitrary (approximate) P-value with subscript m denoting sample size. This could be a first order P-value of the form $P(Y) = 1 - \Phi(T(Y))$ where T is one of the approximate pivots above or a bootstrap P-value. P-values should be uniformly distributed and so ideally the exact type 1 error

$$\Pr(P_m(Y) \leq \alpha; \psi_0, \lambda)$$

should equal α for all λ . In discrete examples, values of λ on the boundary typically correspond to degenerate distributions concentrated on data sets that provided no information. Standard likelihood based P-values tend to be non-significant for these data sets. For instance, in a logistic regression extreme values of the nuisance parameter correspond to a degenerate distribution where all successes or all failures are observed and the SRLR statistic takes the value zero. In this case $\Pr(P_m(Y) \leq \alpha; \psi_0, \lambda)$ typically converges to zero at the extremes, for all reasonable P-values.

The relative error in a P-value will be measured by

$$e_m(\alpha, \lambda) := \Pr(P_m(Y) \leq \alpha; \psi_0, \lambda) / \alpha - 1 \tag{4}$$

Since this depends on λ we will summarise it in two ways. The mean absolute relative error is

$$\bar{e}_m(\alpha) = \int |e_m(\alpha, \lambda)| dG(\lambda)$$

where we will take G to be Lebesgue for a natural parametrisation of $\lambda \in [0, 1]$. We will also look at

$$e_m^*(\alpha) = \left| \sup_{\lambda} e_m(\alpha, \lambda) \right|$$

which measures the worst exaggeration of the significance of the data. This is a one-sided measure in the sense that $\Pr(P_m(Y) \leq \alpha; \psi_0, \lambda)$ could be much smaller than α for some values of λ without $P(Y)$ being harshly judged. This is consistent with Bickel and Doksum (1997), who suggest that the quoted size of a test be maximised with respect to the nuisance parameter. It is also consistent with theory of Rohmel and Mansmann (1999) and Lloyd (2008) who list some strong optimality properties of maximised P-values $P_m^*(Y)$ defined by

$$P_m^*(y) = \sup_{\lambda} \Pr(P_m(Y) \leq P(y); \psi_0, \lambda).$$

A third α -free measure is based on comparing the P-value $P_m(Y)$ with its maximised version $P_m^*(Y)$ over the whole sample space. We calculate the mean value of $|P_m^*(Y) - P_m(Y)|$ with respect to the model $p_m(\cdot; \lambda)$ on Y and the uniform distribution on λ . This measure answers the question "how close on average is the quoted P-value to the significance level we would be forced to quote if we want to guarantee the test."

4 Non-inferiority tests

In a clinical trial comparing a new treatment to a standard treatment, one may want to demonstrate that the new treatment is not practically inferior to the standard treatment. If p_1 is the probability of a positive end-point with the new treatment and p_0 the probability under the old, then we want to demonstrate that p_1 is not too much smaller than p_0 . So-called non-inferiority test decide between the hypotheses

$$\mathcal{H}_0 : p_1/p_0 \leq 1 - \delta \quad \text{versus} \quad \mathcal{H}_1 : p_1/p_0 > 1 - \delta,$$

for some pre-chosen non-inferiority margin δ , often 0.1. Rejecting the null hypothesis means that the new treatment is not practically inferior to the old treatment, and may actually be better. Such tests were first considered by Chan (1998).

The data comprise y_0 responses from n_0 independent individuals with treatment not applied, and y_1 responses from n_1 independent individuals with treatment applied. The parameters of interest are the probabilities of response p_0, p_1 with and without treatment. Denoting the interest parameter by $\psi = \log(p_1/p_0)$, the non-inferiority hypotheses are

$$\mathcal{H}_0 : \psi \leq \psi_0 \quad \text{versus} \quad \mathcal{H}_1 : \psi > \psi_0,$$

where $\psi_0 = \log(1 - \delta)$. Letting the nuisance parameter $\lambda = p_0$ the log-likelihood is

$$\ell(\psi, \lambda; y_0, y_1) = y_1 \psi + t \log \lambda + (n_1 - y_1) \log(1 - \lambda e^\psi) + (n_0 - y_0) \log(1 - \lambda) \quad (5)$$

where $t = y_0 + y_1$ is the total number of successes. The score vector has λ component

$$\frac{\partial \ell}{\partial \lambda} = \frac{t}{p_0} - \frac{(n_1 - y_1)e^\psi}{1 - p_0 e^\psi} - \frac{n_0 - y_0}{p_0}$$

and so the MLE $\hat{p}_{0\psi}$ of p_0 when ψ is fixed is obtained by solving a quadratic whose coefficients turn out to equal

$$c_2 = (n_0 + n_1)e^\psi, c_1 = -(y_1 + n_0) - (y_0 + n_1)e^\psi, c_0 = t.$$

This estimator, in a different form, was first given by Miettinen & Nurminen (1985). The restricted ML estimator of p_1 is $\hat{p}_{1\psi} = e^\psi \hat{p}_{0\psi}$. The so-called likelihood root statistic is

$$T(y_0, y_1; \psi) = \text{sign}(\hat{\psi} - \psi) \left[2 \left\{ \ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi) \right\} \right].$$

The null distribution is approximately standard normal when $\min\{n_0, n_1\}$ diverges and provided (p_0, p_1) are not on the boundary of the unit square.

A second test statistic suggested by Chan (1998) is based on estimating $p_1 - p_0(1 - \delta)$ and comparing it to zero. This generates an approximate pivot

$$T(y_0, y_1; \psi) = \frac{y_1/n_1 - e^\psi y_0/n_0}{\hat{\tau}_\psi}$$

where $\hat{\tau}_\psi^2 = \hat{p}_{1\psi}(1 - \hat{p}_{1\psi})/n_1 + e^{2\psi} \hat{p}_{0\psi}(1 - p_{0\psi})/n_0$ estimates the variance of the numerator under the null.

One-sided P-values based on standard normal approximation to either of these approximate pivotals are said to be first order accurate and suffer errors of $O(m^{-1/2})$. I computed $\Pr(P \leq \alpha; p_0)$ and then summarised the error by the mean with respect to p_0 , denoted $\hat{e}_m(\alpha)$ and also maximising with respect to p_0 , denoted $e^*(\alpha)$. The values of p_0 were an even grid of 101 points between 0 and 1. The left panel of Figure 1 is of $\Pr(P \leq 0.025; p_0)$ for sample sizes (175, 75) and P equal to the first order P-value based on the SRLR statistic (line) and P equal to the bootstrap P-value based also on the SRLR statistic (bold). The mean absolute relative errors are respectively 0.067 and 0.054 and the relative errors of the supremum are respectively 0.246 and 0.003.

Sample sizes (n_0, n_1) were scaled up by a factor of m from 1 to 225 for the first order P-value and from 1 to 50 for the bootstrap P-value (because of computational limits). The centre panel of Figure 1 shows a plot of $\bar{e}_m(0.025)$ and $e^*(0.025)$ versus m for the LR statistic on the log-log-scale. The right panel is for the bootstrap P-value. It is noteworthy that, for the standard LR statistic, the maximum error is much worse

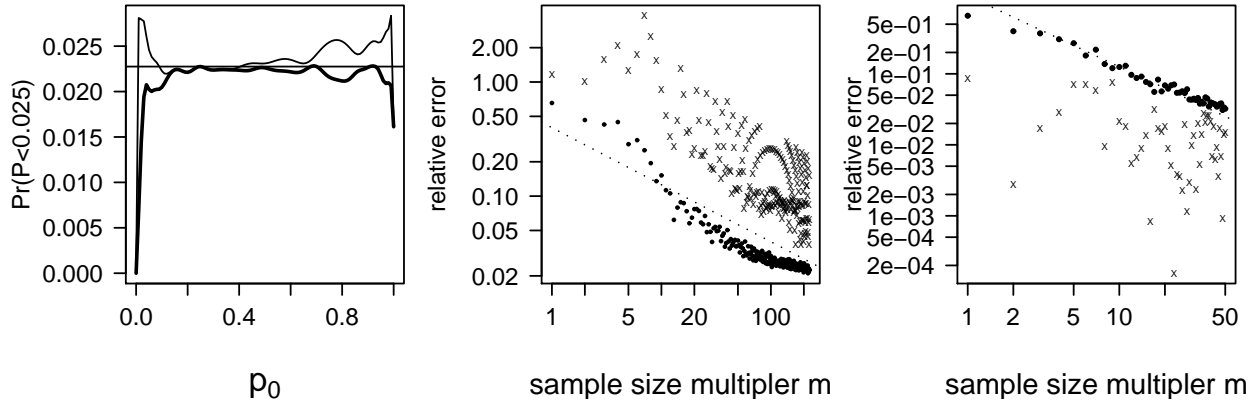


Figure 1: *Left.* $\Pr(P \leq 0.025; p_0)$ versus p_0 for $(n_0, n_1) = (175, 75)$. *Centre.* Log-log plot of $\bar{e}_m(0.025)$ (circle) and $e_m^*(0.025)$ (cross) versus m , for SRLR statistic. Line has slope $-1/2$. *Right.* Log-log plot of $\bar{e}_m(0.025)$ (circle) and $e_m^*(0.025)$ (cross) versus m for bootstrap SRLR statistic. Line has slope -1 .

than the mean error, decreasing from around 200% to around 10% when $m = 225$ which is for a table with margins $(1575, 675)$. Looking at the error measures $\bar{e}_m(0.025)$ represented by circles, it appears that the accuracy of the standard P-value improves at the rate $m^{-1/2}$ whereas for the bootstrap P-value the rate appears to be m^{-1} .

To investigate the asymptotic rates, I fitted a model of the form

$$\bar{e}_m(\alpha) = b_1 m^{-1/2} + b_2 m^{-1} + b_3 m^{-3/2}$$

and estimated b_1, b_2, b_3 by weighted least squares. A small value of b_1/b_2 indicates that the $m^{-1/2}$ term is vanishing and a small value for b_2/b_3 would indicate that the m^{-1} term is vanishing. I also estimated the single best rate γ from a regression of $\log \bar{e}_m(\alpha) = c + \gamma \log m$. The results are in Table 2. The results are consistent with the assertion that $\bar{e}_m(\alpha)$ is $O(m^{-1})$.

Results for $e_m^*(\alpha)$ are less clear. It is apparent from the plots, and also the results in Table 4, that dependence on m is very erratic. It does not appear that for bootstrap P-values $e_m^*(\alpha)$ decreases at the rate m^{-1} as is the case for $\bar{e}_m(\alpha)$. Notwithstanding the rates, the actual value of the errors is much smaller for bootstrap P-values than ordinary P-values. To demonstrate this, I estimated $\hat{e}_m^*(\alpha)$ for $m = 20$ from the first regression model. In the first row, we can see that the maximum error in the P-value is 130% (of 0.025) and this reduces to 4% for the bootstrap P-value.

Table 2: **Mean absolute error rates of standard and bootstrap P-values.** For each experiment, three figures described in the text summarise the asymptotics. From a regression of the form $\bar{e}_m(\alpha) = b_1 m^{-1/2} + b_2 m^{-1} + b_3 m^{-3/2}$ the ratio b_1/b_2 and b_2/b_3 indicates the relative importance of the $m^{-1/2}/m^{-1}$ terms. The third figure is the estimated rate from a log-log regression of $\bar{e}_m(\alpha)$ on m .

Experiment			First order P-values			Bootstrap P-values		
(n_1, n_2)	T	α	b_1/b_2	b_2/b_3	rate γ	b_1/b_2	b_2/b_3	rate γ
(7,3)	chan	.025	6.09	0.41	-0.58	0.06	1.99	-1.09
(7,3)	chan	.05	0.69	0.61	-0.52	0.10	2.09	-0.95
(7,3)	LR	.025	0.64	1.88	-0.47	0.01	1.72	-1.01
(7,3)	LR	.05	1.14	0.25	-0.50	0.09	2.08	-0.96
(4,6)	chan	.025	0.42	2.81	-0.68	0.06	2.21	-0.98
(4,6)	chan	.05	0.66	11.32	-0.58	0.02	1.98	-1.02
(4,6)	LR	.025	0.26	17.88	-0.72	0.01	1.80	-0.98
(4,6)	LR	.05	0.77	11.81	-0.58	0.04	1.94	-0.99

5 Logistic regression

Consider binomial data (y_1, \dots, y_d) with binomial denominators (n_1, \dots, n_d) and probability parameters π_i following the logistic regression

$$\pi_i = \frac{\lambda e^{\psi x_i}}{1 - \lambda + \lambda e^{\psi x_i}}.$$

The parameter λ is the expit transform of the usual intercept parameter. and represents the value of π_i when $x_i = 0$ and ranges over $[0, 1]$. Tests of ψ are often made conditional on the sufficient statistic $\sum y_i$ for λ but we do not address such tests here. Rather, we are interested in the unconditional properties of some standard tests and the rate at which errors in quoted tail probabilities reduce to zero. It was noted in the introduction that DiCiccio and Young (2008) claim that bootstrap test, like so-called r^* based tests, are approximations to the conditional tests, while Young and Lee (2005) show that bootstrap tests have accurate unconditional properties.

We will investigate two statistics. The first is the signed root likelihood ratio statistics which can be expressed as the signed root of

$$\text{LR} = 2 \sum_{i=1}^d \hat{y}_i \log \left(\frac{\hat{y}_i}{\hat{y}_{i0}} \right) + 2 \sum_{i=1}^d (n_i - \hat{y}_i) \log \left(\frac{n_i - \hat{y}_i}{\hat{n}_i - \hat{y}_{i0}} \right)$$

with \hat{y}_i and \hat{y}_{i0} denoting fitted values under the unrestricted/restricted logistic regression models respectively. To evaluate accuracy, this statistic must be computed for all

Table 3: **Worst Error rates of standard and estimated P-values.** For each experiment, four figures describe the performance. The first three figures are ratios b_1/b_2 , b_2/b_3 and rate γ for regressions $|e_m^*(\alpha)|$. The fourth figure is the extrapolated worst error rate when $m = 20$.

Experiment			First order P-values				Estimated P-values			
(n_1, n_2)	T	α	b_1/b_2	b_2/b_3	rate	\hat{e}_{20}^*	b_1/b_2	b_2/b_3	rate	\hat{e}_{20}^*
(7,3)	chan	.025	1.05	0.38	-0.90	1.30	0.63	1.26	-0.50	0.04
(7,3)	chan	.05	0.36	0.97	-0.43	0.30	0.90	1.72	-0.58	0.03
(7,3)	LR	.025	0.05	1.07	-0.62	0.87	1.72	1.02	-0.61	0.03
(7,3)	LR	.05	0.11	1.26	-0.59	0.71	0.52	1.30	-0.30	0.03
(4,6)	chan	.025	0.08	1.59	-1.02	0.16	0.55	1.35	-0.42	0.11
(4,6)	chan	.05	0.04	0.96	-0.52	0.14	0.38	1.20	-0.02	0.06
(4,6)	LR	.025	0.12	0.82	-1.10	1.89	0.71	0.88	-0.30	0.07
(4,6)	LR	.05	2.42	1.39	-0.51	0.61	0.43	1.05	-0.25	0.05

possible data sets, which become computationally difficult. Therefore, we also investigate an alternative statistic with identical form to above but with \hat{y}_i and \hat{y}_{i0} estimated by weighted least squares on the empirical logits, with one iteration on the weights. These statistics can be computed on the whole sample space simultaneously since they are just linear transformations.

We consider testing $\psi > 0$ for sample sizes (3, 4, 5) scaled up by a multiplier m . The computational burden becomes extreme and increases as m^6 . With $m = 10$ there are $N = 127551$ points in the sample space and computing all bootstrap P-values required $O(N^2)$ computations. Figure 2 displays the accuracy measure $\bar{e}_m(\alpha)$ for $m = 1, \dots, 10$ for the ordinary and bootstrap P-values based on the SRLR statistic (left panel) and WLS statistic (right panel). It is apparent that the bootstrap P-value has lower levels of error and the rate of decrease with m appears larger.

Table 4 presents results on the asymptotic rates implied by these plots. There is again evidence that the $m^{-1/2}$ term is vanishing but no evidence that the m^{-1} term is vanishing. In each case, the best estimate of the slope for the bootstrap P-values is somewhat less than -1, but this appears to be a limitation of the computations.

As with the non-inferiority test, maximum error tells a different story to mean error. In Figure 3 there is no clear tendency for error to decrease with sample size up to $m=10$. What is clear is that the bootstrap P-value has error an order of magnitude smaller than standard P-values. The right plot shows the actual significance $\Pr(P \leq 0.05; p_0)$ versus p_0 for the LR statistics and bootstrap LR statistics for $m = 5$.

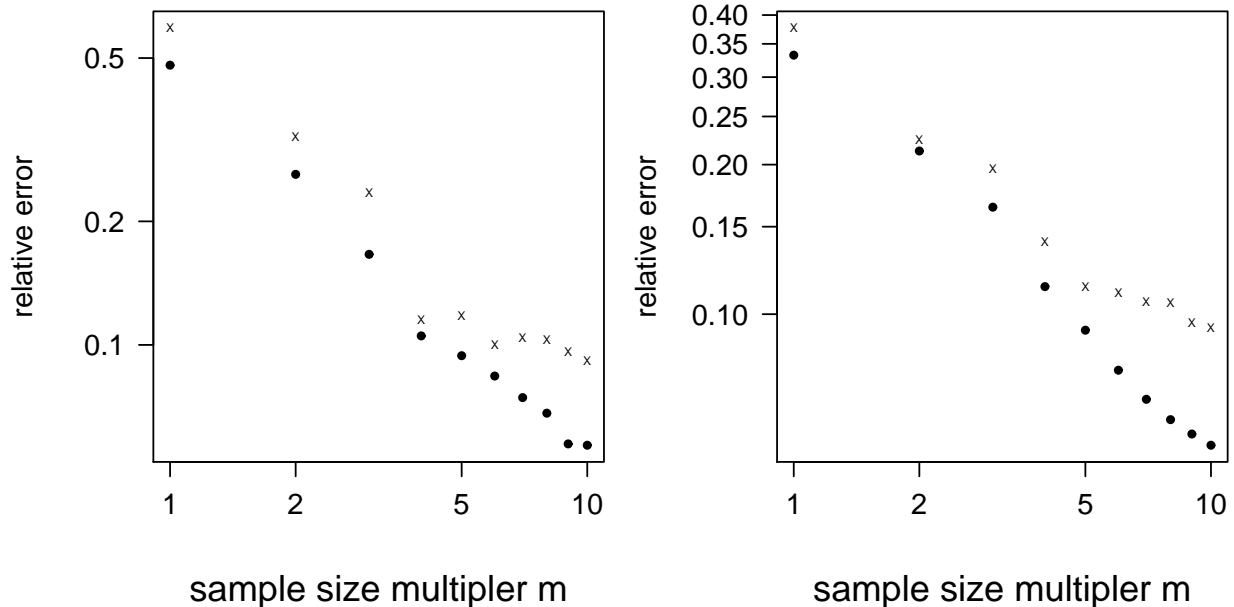


Figure 2: *Left.* Log-log plot of $\bar{e}_m(0.05)$ for P-value based on LR statistic (cross) and bootstrap LR statistic (circle) *Right.* Log-log plot of $\bar{e}_m(0.05)$ for P-value based on WLS statistic (cross) and bootstrap WLS statistic (circle)

Why does it work?

While it appears that average error rates of bootstrap P-values for discrete models decrease at rate $O(m^{-1})$, the more pertinent conclusion from our numerical study is that worst error is an order of magnitude smaller than naive P-values and is not very dependent on sample size. In contrast to the examples investigated by Lee and Young (2005), the gains are far from slight. These gains do not appear to be an asymptotic phenomenon. It is presumably due to \hat{P} ranking points in the sample space more correctly in terms of their hostility to the null. It is certainly plausible that the best estimate of the exact significance of a data point, namely \hat{P} , is likely to give a good data based ranking of significance.

To put a little more meat on this argument, suppose that there exists a P-value $U(Y)$ which is very close to uniformly distributed for all λ . Let u_j denote the j th largest value of $U(Y)$ and y_j the corresponding data point. The benchmark P-value $U(Y)$ has the property that

$$\Pr\{U(Y) \leq u_j\} = R(u_j, \lambda) \tag{6}$$

where $R(y, \lambda) - y = e(y, \lambda)$ is uniformly small for all λ .

Table 4: **Mean absolute error rates of standard and bootstrap P-values.** The model is a logistic regression with sample sizes (3, 4, 5) scaled up by a factor $m=1, \dots, 10$. For each experiment, three figures summarise the asymptotics. From a regression of the form $\bar{e}_m(\alpha) = b_1 m^{-1/2} + b_2 m^{-1} + b_3 m^{-3/2}$ the ratio b_1/b_2 and b_2/b_3 indicates the relative importance of the $m^{-1/2}$ and m^{-1} terms respectively. The third figure is the estimated rate from a log-log regression of $\bar{e}_m(\alpha)$ on m .

Experiment		First order P-values			Bootstrap P-values		
T	α	b_1/b_2	b_2/b_3	rate γ	b_1/b_2	b_2/b_3	rate γ
lr	05	3.69	0.15	-0.51	0.13	4.18	-0.82
lr	025	0.79	0.90	-0.55	0.08	1.50	-0.81
wls	05	3.02	0.54	-0.49	0.05	2.05	-0.87
wls	025	0.35	1.23	-0.46	0.05	2.03	-0.85

To investigate the effect of the bootstrap adjustment, we look at a statistic $Q(Y)$ which does not share the good properties of $U(Y)$ and show that the bootstrap/estimation step transforms $Q(Y)$ towards $U(Y)$. We look at two distinct deficiencies in $Q(Y)$ that may be corrected. First, $Q(Y)$ may be close to pivotal but simply not have the uniform distribution. In this case, it can be transformed to near uniformity and we show that the bootstrap step achieves this. Second, $Q(Y)$ may differ from $U(Y)$ by imposing an incorrect ordering on the sample space. We look at the case where sets has been mis-ranked and show that the bootstrap step returns the data set to its correct rank without unduly disturbing the ranking of other data sets.

Let $Q(Y)$ be another statistic with atoms $q_j = Q(y_j)$. To express the assumption that $Q(Y)$ is approximately pivotal but not uniform we suppose that

$$\Pr\{Q(Y) \leq q_j\} = R(h(q_j), \lambda) = h(q_j) + e(h(q_j), \lambda).$$

for a monotone function h . Since this is identical to $\Pr\{h(Q(Y)) \leq h(q_j)\}$, it follows that $U(Y) := h(Q(Y))$ satisfies (6) with $u_j = h(q_j)$. The bootstrap P-value based on Q is

$$\hat{Q}(y_j) = R(h(q_j), \hat{\lambda}(y_j)) = h(q_j) + e(h(q_j), \hat{\lambda}(y_j))$$

and so

$$\hat{Q}(Y) = h(Q(Y)) + e(h(Q(Y)), \hat{\lambda}(Y)) = U(Y) + e(U(Y), \hat{\lambda}(Y)).$$

The conclusion then is that for an approximately pivotal but highly non-inform variable $Q(Y)$, the bootstrap step approximately applies the correct transformation h towards approximate uniformity.

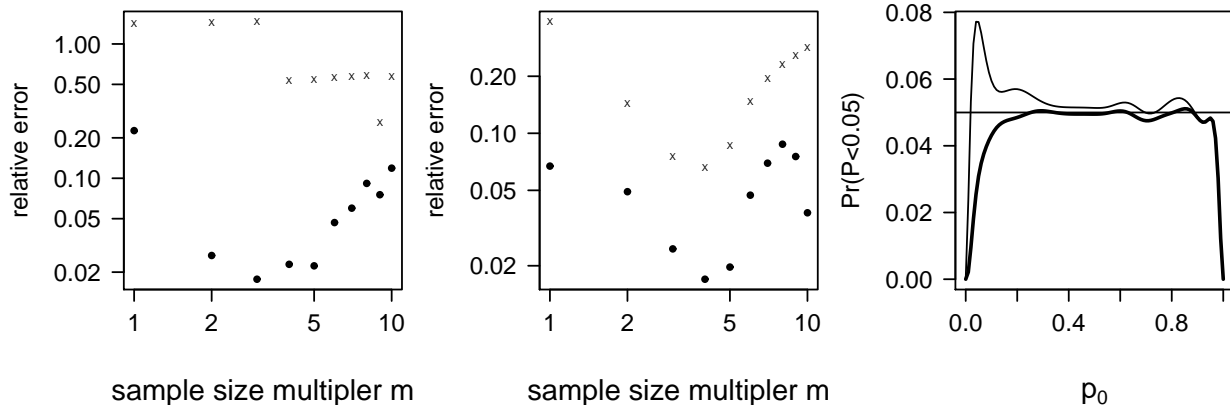


Figure 3: *Left.* Log-log plot of $e_m^*(0.05)$ for P-value based on LR statistic (cross) and bootstrap LR statistic (circle) *Centre.* Log-log plot of $\bar{e}_m(0.05)$ for P-value based on WLS statistic (cross) and bootstrap WLS statistic (circle). *Right.* $\Pr(P \leq 0.05; p_0)$ versus p_0 for the LR statistics with sample sizes (15, 20, 25).

To look at the issue of ordering, it is worth first noting that for any statistic $Q(Y)$ with atoms y_j , the null tail probabilities are a sum of terms $\pi(y_i; \lambda)$ each of which are highly non-pivotal and are maximised at the restricted ML estimator $\hat{\lambda}_i$. Denote by $\hat{\pi}_{i,k}$ the value of $\pi(y_i; \lambda)$ at $\hat{\lambda}_k$. When $i = k$ these will be much larger than when $i \neq k$, which will be very close to zero when $\hat{\lambda}_i$ is far from $\hat{\lambda}_k$. In all that follows, it is essential that $\hat{\lambda}_k$ be the restricted ML estimator, rather than the unrestricted.

Now let $Q(Y)$ be a statistic that is identical to $U(Y)$ for all data sets except a particular data point y_m . Suppose that $u_L < Q(y_m) < u_{L+1}$ where $m > L + 1$. So the statistic Q mistakenly ranks y_m as more significant than data points y_{L+1}, \dots, y_{m-1} . The hope is that bootstrap adjustment will restore y_m to its rightful ordering, in other words we hope that $\hat{q}_m > \hat{q}_j$ for $j = L + 1, \dots, m - 1$.

For most data sets, the mis-ranking of y_m has no effect on inference because for $j \leq L$ or $j \geq m + 1$,

$$\Pr\{Q(Y) \leq q_j\} = \Pr\{U(Y) \leq u_j\}$$

and so $\hat{Q}(y_j) \approx q_j = u_j$. On the other hand, for $j = L + 1, \dots, m - 1$

$$\Pr\{Q(Y) \leq q_j\} = \Pr\{U(Y) \leq u_j\} + \pi(u_m; \lambda) = R(u_j; \lambda) + \pi(u_m; \lambda).$$

All of these tails functions are highly non-pivotal because they comprise a flat function $R(u_j; \lambda)$ with $\pi(u_m; \lambda)$ added. The bootstrap values are

$$\hat{q}_j = \hat{u}_{j,j} + \hat{\pi}_{m,j}$$

where $\hat{u}_{j,j} = R(u_j; \hat{\lambda}_j)$ which is approximately equal to u_j by supposition. For $j = m$,

$$\Pr\{Q(Y) \leq q_m\} = \Pr\{U(Y) \leq u_L\} + \pi(u_m; \lambda) = R(u_L; \lambda) + \pi(u_m; \lambda)$$

is also non-pivotal and

$$\hat{q}_m \approx u_L + \hat{\pi}_{m,m} = \hat{q}_j + \hat{\pi}_{m,m} - \hat{\pi}_{m,j} - (u_j - u_L).$$

Firstly, provided that $\hat{\lambda}_j$ is not close to $\hat{\lambda}_m$, it follows that

$$\hat{q}_m > \hat{q}_j \Leftrightarrow \hat{\pi}_{m,m} > u_j - u_L.$$

So y_m will be ranked after y_j so long as the maximum probability of y_m is larger than the partial sum of probabilities of y_{L+1} up to y_j . Is it true that $\hat{\lambda}_j$ is not close to $\hat{\lambda}_m$? If it were, then we would have two data sets which indicated a similar value of λ , where $\pi(y_m; \lambda)$ was larger than $\pi(y_j; \lambda)$, and yet y_m was being ranked as more significant than y_j , which would be a grossly illogical ordering.

To reveal more of why bootstrap works, it is necessary to make plausible assumptions about how the sample space is initially ranked. Indeed, the bootstrap does not work for a completely random ranking of the sample space. The right panel of Figure 4 displays a typical tail set based on a test statistic which measures statistical departure from the null, as indicated by the broken line. While this panel represents a particular tail set for the non-inferiority example, it should be treated as generic. The horizontal axis lists data sets in order of the values of $\hat{\lambda}_j$ (which could be multi-dimensional). The vertical axis is the remaining degree of freedom of the data, chosen such that larger values indicate more departure from the null. The solid points are on the boundary of the tail set where the test statistic T approximately equals some fixed value, for instance 2. For one of these points T will equal 2, and for the others it will be less than 2 by small amounts which are determined by the vagaries of discreteness and are hard to predict. So the ranking of points on the boundary will be basically random.

If T is related to a likelihood ratio statistic then the probabilities of points on the boundary will be similar in magnitude. As we move further into the tail set the probabilities become smaller rather quickly. Imagine that the errant point y_m is the point (2, 6) represented by the larger circle. This point will have larger probability than the original boundary points. The point y_L is the point (2, 7) just above it, and the remaining points on the boundary are y_{L+1}, \dots, y_{m-1} in the earlier exposition.

The probability profiles of the boundary points $\pi(y_j; \lambda)$ are represented as solid lines in the right panel of Figure 4. Each of these is maximised at $\hat{\lambda}_j$ and spread evenly

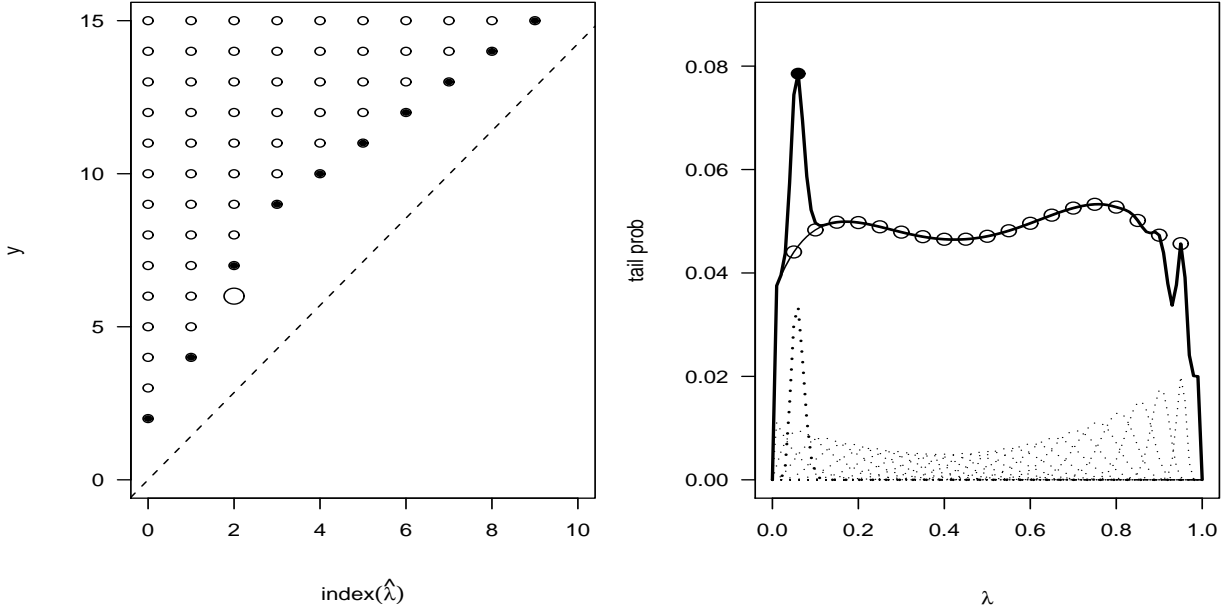


Figure 4: *Left.* Generic tail sets. Boundary points have similar null probabilities $\hat{\pi}_{j,j}$ and similar values for $T(y_j)$. Larger circle is the errant point y_m . *Right.* Generic tail probability. Errant point's probability profile $\pi(y_m; \lambda)$ is bold dotted. Profiles of other points are smaller and solid. Overall profile of tailset contains spike at $\hat{\lambda}_m$.

across the range of λ so that when they are accumulated the overall tail probability is reasonably pivotal. Inclusion of the errant point y_m however adds a larger additional spike of probability and results in a non-pivotal tail probability. We can now argue that the bootstrap will restore y_m to its correct ranking, namely as less significant than any of y_{L+1}, \dots, y_{m-1} .

The value of \hat{q}_m is marked as a filled circle and is the supremum of this plot. This is because the tail set for y_m includes both y_m and y_L both of which are maximised at $\hat{\lambda}_m$. For the points j on the boundary, the tail set includes y_L, y_M, y_j itself and all other points on the boundary judged more significant than y_j . Near $\lambda = \hat{\lambda}_j$, the tail probability will be very close to the relevant part of this bold curve. The bootstrap value \hat{q}_j will be this bold curve evaluated at $\hat{\lambda}_j$ and depicted as open circles. These are all smaller than \hat{q}_j so the ranking of y_j is restored to being less significant than any of these boundary points. The bootstrap ranking of the points on the boundary will mimic the ranking of $\hat{\pi}_{j,j}$ rather than the essentially random ranking based on the original test statistic T .

It is worth point out that if y_m were ranked more significant than y_L , then \hat{q}_m will

not equal the supremum of the plot, because y_L will not be in the tail set. Rather it will equal this maximum minus $\hat{\pi}_{L,L}$ which could easily be smaller than \hat{q}_j on the boundary. So, for a sufficiently illogical ordering, the bootstrap cannot restore y_m to its correct position. Such mis-ranking will not occur for standard test statistics. Rather, it is precisely points such as the one illustrated that are problematic in practice and which bootstrap can correct.

6 Conclusion

Bootstrap P-values differ from their theoretical uniform distribution by $O(m^{-1})$ according to the numerical evidence of this study. However, worst error is much less affected by sample size and the advantages of bootstrap P-values are much clearer for this measure. We have explained how bootstrap P-values automatically correct for test statistics which are pivotal but have been calibrated against the wrong distribution. We have further shown that for P-values based on likelihood statistics, bootstrap will correct the mis-rankings of the sample space induced by the test statistic.

Computing bootstrap P-values can be implemented by simulating the model under the null estimate $(\psi_0, \hat{\lambda}_0)$ however this is likely to be inefficient for smaller tail probabilities. The author is currently developing algorithms which simulate under the unrestricted estimate and corrects the estimate using importance sampling ideas. This also allows the significance profile $\Pi(t, \lambda)$ to be conveniently approximated.

References.

- Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83**, 687-697.
- Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics*. Holden-Day, Oakland.
- Brazzale A.R., Davison A.C. and Reid N. (2007) *Applied asymptotics: case studies in small sample statistics*. CUP, Cambridge.
- DiCiccio, T.J., Martin, M.A. and Stern, S.E. (2001) Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canad. J. Statist* **29**, 67–76.
- DiCiccio, T.J. and Stern, S.E. (1994) Constructing approximate standard normal pivots from signed roots of adjusted likelihood ratio statistics. *Scand. J. Statist* **21**, 447-460.
- Fraser, D.A.S and Rousseau, J. (2008) Studentization and deriving accurate P-values. to appear in *Biometrika*.
- Lee S.M.S. and Young, G.A. (2005) parametric bootstrapping with nuisance parameters. *Stat. prob. letters* **71**, 143–153.
- Lloyd, C.J. (2008) Exact P-values for discrete models obtained by estimation and maximisation. To appear in *Austral. and New Zealand J. Statist.* **50**.
- Miettinen, O. and Nurminen, M. (1985) Comparative analysis of two rates. *Statistics in Medicine* **4**, 213-226.
- Rohmel J, Mansmann U. (1999) Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority or superiority. *Biometrical Journal* **41**, 149–170